NeuroImage

Separate encoding of model based and model free valuations in the human brain

Ulrik R. Beierholm[1], Cedric Anen[2], Steven Quartz[2,3], Peter Bossaerts[2,3,4]

[1]*Gatsby Computational Neuroscience Unit, UCL, London, UK*

[2]*Computation and Neural Systems, Caltech, Pasadena, CA-91125, USA*

[3]*Division of Humanities and Social Sciences, Caltech, Pasadena, CA-91125, USA*

[4]*Laboratory for Decision Making under Uncertainty, EPFL, Lausanne, Switzerland*

Corresponding author:

Ulrik Beierholm

Gatsby Unit, 17 Queen Square

London, WC1N 3AR

UK

Phone: +44 (0) 20 7679 1185

Email: beierh@gatsby.ucl.ac.uk

Keywords: Dual systems, reinforcement learning, Bayesian inference, striatum, vmPFC, decision making

Abstract

Behavioral studies have long shown that humans solve problems in two ways, one intuitive and fast (System 1, model-free), and the other reflective and slow (System 2, model-based). The neurobiological basis of dual-process problem-solving remains unknown due to challenges of separating activation in concurrent systems. We present a novel neuroeconomic task that predicts distinct subjective valuation and updating signals corresponding to these two systems. We found two concurrent value signals in human prefrontal cortex: a System 1 model-free reinforcement signal and a System 2 model-based Bayesian signal. We also found a System 1 updating signal in striatal areas and a System 2 updating signal in lateral prefrontal cortex. Further, signals in prefrontal cortex preceded choices that are optimal according to either updating principle, while signals in anterior cingulate cortex and globus pallidus preceded deviations from optimal choice for reinforcement learning. These deviations tended to occur when uncertainty regarding optimal values was highest, suggesting that disagreement between dual systems is mediated by uncertainty rather than conflict, confirming recent theoretical proposals.

Introduction

The notion that cognitive processes can be partitioned into two main categories, intuition and reason, is as old as the study of human thinking. In contemporary form, this distinction underlies various dual system models of judgment and decision making (Evans, 2008). According to these models, choice can be determined either through a fast and reflexive process ("System 1") or through a more reflective and slow process ("System 2"). Pointed examples in the context of problem solving involving probabilities have System 1 generate quick answers based on substitution of the given problem by another one for which the correct answer is readily available (Kahneman and Frederick, 2002).

Behaviorally, the evidence for the existence of a dual system appears to be clear (Evans, 2008; Fermin et al., 2010; Sloman, 1996). However, its neurobiological foundations have not been established. Recent evidence of activation relating to distinct regions in the human brain depending on context ,(De Martino et al., 2006; Tzieropoulos et al., 2011) is based on inter-subject comparisons. While these findings are consistent with dual system theory (Bossaerts et al., 2008; Kahneman and Frederick, 2007), they leave open the possibility that different subjects engage unique albeit group-specific decision making modules. Different patterns of activations across contexts (Hsu et al., 2005; McClure et al., 2004) likewise have been deemed insufficient evidence in favor of dual-system theory because they could merely reflect divergence in measurement of relevant components, and not fundamental differences in the way these components are integrated to determine choice (Kable and Glimcher, 2007; Levy et al., 2010). Ideally we would like to find evidence for the concurrent processing of the two systems, i.e. the presence of two sets of signals both contributing to the final decision. Naturally this introduces a need for adjudication, for those inevitable situations where the two systems disagree. Theories of how this arbitration may take place have emerged but are as yet untested (Daw et al., 2005; De Neys and Glumicic, 2008; Kerns et al., 2004).

Recent neuroeconomic insights provide a novel approach to testing whether dual systems underlie human choice. A basic principle of economics is that choice is based on subjective valuation and that decision makers choose the option that maximizes their subjective utility. While this has long been interpreted as an “as if” procedure (the decision maker chooses as if she is maximizing a utility function), recent evidence in the neuroeconomics literature supports the idea that this is to be taken literally. Specifically, signals have been identified in the human (Plassmann et al., 2007; Tobler et al., 2007) and nonhuman primate (Padoa-Schioppa and

Assad, 2006; Platt and Glimcher, 1999; Tremblay and Schultz, 1999) brain that encode the subjective valuations revealed through choice, and these signals predict future choice (Tobler et al., 2007). Further, related studies established the existence of an update signal alongside the valuation signal, thus allowing the study of the process that drives changes in these valuations (McClure et al., 2003; O'Doherty et al., 2003).

A new neurobiological hypothesis emerges when we confront dual-system theory with these neuroeconomic insights. Specifically, if choice were based on the evaluation of the outcomes of two systems, it would predict the existence of two concurrent subjective value signals in the brain, a System 1 value signal and a System 2 value signal. Work done by Daw et al. 2005 directly provides a way to compare the systems by identifying System 1 with a habitual model-free reinforcement learning system, and the rule based system with a near-optimal model-based (e.g. Bayesian) system. Further, because the two systems at times generate conflicting choice, these two value signals would not always coincide and would require a means to adjudicate between them. Here, we test this new hypothesis by searching for the existence of such concurrent subjective value signals, whether such signals predict subject choice, and the conditions under which one or the other signal guides subject choice (adjudication).

Methods

Twenty three subjects, Caltech students as well as subjects recruited by online advertising, participated in the experiment. All subjects were instructed about the general purpose of the experiment and signed a consent form as approved by the Caltech IRB. Subjects were given written instructions for the experiment and were tested briefly to make sure they understood the mechanics of the game.

<INSERT FIGURE 1 AROUND HERE>

Stimuli and game

Subjects were placed inside the MRI-scanner and all stimuli were presented on MRI compatible goggles and generated using the Psychophysics toolbox (Brainard, 1997). Subjects were visually presented with three doors and instructed to choose the order of the doors at their own pace. The average response time was 1.9 s.

After 6-8 seconds the location of the money would be revealed behind one of the doors and subjects would be rewarded according to the following: $.50 if the money was behind their first choice, $0 if behind their second choice and $-0.50 if the money was behind the third choice. They therefore had an incentive to try to order the doors according to the likelihood of the money appearing behind that door.

As part of the instructions they had been informed that the sequence of locations for the money had been chosen randomly before the experiment, and was therefore independent of their choices, and that money was equally likely to appear behind a door in each round, i.e. the likelihood did not change over time.

Each subject played three sessions of 40 rounds with different distributions over the doors in each session. The money symbol had the following chance of appearing behind the doors (left, center, right): [0.1, 0.3, 0.6] in the first session, [0.4, 0.4, 0.2] in the second session, and [0.2, 0.2, 0.6] in the third session. The beginning and end of each session were clearly indicated and subjects were explicitly instructed to ignore anything they learned in previous sessions regarding the distribution of money. For the purposes of the modeling of the learning we assume that the learning models are reset at the beginning of each session.

Subjects made between $0 and $18 during the three sessions and were also paid a show up fee of $20.

System 1 model

The behavioral data from each subject, i.e. the order of the doors in each round, was used to fit the learning rate, α, in a simple reinforcement learning (RL) model. According to this model in each round, t, the expected value $V_{j,t}$ of choosing a combination of doors (e.g. left, right, center) gets updated at the time of the reveal screen according to:

$V_{j,t}=V_{j,t-1}+ \alpha * (Reward_t - V_{j,t-1})$.

This is the Rescorla-Wagner updating rule (Rescorla and Wagner, 1972) which is a special case of the Q-learning algorithm (Sutton and Barto, 1998). Only the chosen order of doors, $V_{j,t}$ get updated in a round. The expected value (reward) for a certain round, t, is given as max($V_{j,t}$). We assumed that the likelihood of subjects' choosing a certain combination of doors, j, is given by a softmax function over the values $V_{j,t}$:

$$likelihood_{j,t} = \exp(\beta * V_{j,t}) / \sum_{k} \exp(\beta * V_{k,t})$$

We used the maximum likelihood (ML) of the data given the model to fit the learning rate, α, and the softmax steepness, β, to the data. The expected values in the model of the subjects' choices were therefore dependent on the choices themselves as well as the individual learning rate. The steepness of the softmax, represented by β, captures any randomness in the subject behavior; a high beta signifies very little deviation from a model, while a low value means the subject performs seemingly random. The fitted learning rate was only used for behavioral model comparison, for the fMRI results a fixed value of 0.4 was used in order to regularize the results and avoid overfitting (similar to previous modeling of fMRI data (Daw et al., 2006; Gläscher et al., 2010)). The learning rate was chosen arbitrarily, but as behavior should reflect both a fixed system 1 learning rate and an effectively decreasing system 2 learning rate we would expect the true system 1 learning rate to be larger than the fitted value. The results presented here are not dependent on the exact learning rate (see SOM).

When fitting the learning rate to the subject data, we found a negative learning rate for five subjects, indicating they were not performing the task correctly, instead following a strategy closer to the Gamblers Fallacy (Tversky and Kahneman, 1974). We therefore chose to disregard these subjects for the purposes of this study, leaving 18 subjects.

As a measure of the subjective uncertainty of System 1 we used the exponentially smoothed average of past squared prediction errors:

$$U_{1,t} = \sum_{s=1}^{\infty} \lambda^{s} (\text{Reward}_{t-s} - \max_{j}(V_{j,t-s-1}))^{2}$$

where for the fMRI analysis we set the steepness of the decay equal to the learning rate, λ=0.4. While the choice of value for this parameter is arbitrary it seems reasonable to expect it to operate on the same time scale as the learning rate.

System 2 model

An optimal Bayesian observer (System 2) was also applied to the data using a Dirichlet prior updated by a multinomial likelihood function. At each round, *t*, the posterior distribution is also a Dirichlet distribution with mean, the probability of the reward being behind door *i*, given by

Encoding of model based and model free valuations

$$p_{i,t} = \frac{n_i + 1}{\sum_k n_k + 1}$$

where n_i is the previous number of times the reward has been observed behind door *i*. Given $p_{i,t}$, the Bayesian expected value of the game is given by

$$< R_t >= \sum_k p_{k,t} r_k$$

where *k* is the chosen order of the doors and r_1=\$-0.50, r_2=\$0 and r_3=\$0.50. Hence the optimal thing to do is to order the doors by their posterior probability of hiding the money, so that $< R_t >= \$0.50*(\max_i \left(p_{i,t} \right) - \min_i \left(p_{i,t} \right))$.

The expected value in the Bayesian model was therefore only dependent on the history of appearances of the money, independent of the subjects' subsequent choices and hence identical across subjects. For the purpose of comparing the models, we also used a variant of the softmax function to calculate the likelihood of the subjects' choices:

$$likelihood_{j,t} = (p_{j,t})^{\gamma} / \sum_k (p_{k,t})^{\gamma}$$

Notice that the posterior probability is bound between 0 and 1 (unlike the value in System 1), hence we use the log of the probability for comparison in the softmax, leading to the equation above. We used the maximum likelihood (ML) of the data given the model to fit the steepness parameter, γ, individually for each subject.

As a measure of the subjective uncertainty of System 2 we used the full entropy of the distribution, taking into account the uncertainty of the Dirichlet distribution $P(p_t \mid n_{1,t}, n_{2,t}, n_{3,t})$:

$$U_{2,t} = \int H(p_t) P(p_t \mid n_{1,t}, n_{2,t}, n_{3,t}) dp_t$$

where the entropy $H(p_t) = -\sum_{i=1}^{3} p_{i,t} \log(p_{i,t})$.

fMRI acquisition

Subjects were scanned in a Siemens Trio 3T using a T1-weighted MPRAGE anatomical sequence (256 x 256 matrix, 176 1mm sagittal slices) and subsequently while performing the task. The

sequence used for the acquisition of the BOLD images was a T2*-weighted PACE EPI (TR = 2000 ms, TE = 30 ms, 64 x 64 matrix, 3 x 3 mm^2, 30 3mm slices, no gap). The scanning session lasted 35-40 minutes (~1100 scans each).

Data was processed and analyzed using Brainvoyager version 1.8 (Brain Innovation, Netherlands) and MATLAB (Mathworks, MA). Preprocessing included motion correction, spatial smoothing (Gaussian half-width 6 mm), slice timing correction, high pass filtering and normalization to Talairach space.

Regressors

A general linear model (GLM) was calculated for each subject using block regressors and parametric regressors. Block regressors were used for the time of the choice screen, the time of the reward screen and the pre-decision period. The parametric regressors were: System 1 (reinforcement learning) value of the game for the pre-decision period max($V_{j,t}$), System 2 (Bayesian) value of the game for the pre-decision period $< R_t >$, System 1 learning error *($Reward_t$ - $V_{chosen,t}$)* at the reward screen and System 2 updating (<R_t> -<R_{t-1}>) at the reward screen. In total 3 block regressors and 4 parametric regressors were put into the same GLM with results presented in Figure 2 and 3. The block regressors represent activity non-specific to the learning, e.g. visual activity, while the parametric regressors track the learning related activity. Notice that the Value regressors are at the same time period and thus have to share the variance in the data.

Separate GLMs were created for Figure 4a-b) based purely on block regressors, with a regressor covering the choice screen, reward screen and the time period 0-4 seconds before the choice screen, aligned to end at the onset of the choice screen. For the pre-choice periods a block was either assigned as 'choice correct' or 'choice incorrect' based on the choice during the following choice screen. A separate GLM was created for System 1 and System 2.

The regressors were ortho-normalized and a random effects analysis was done on the parametric regressors using Brainvoyager (Brain Innovation B.V., Netherlands), testing if the activation was significant on the subject group as a whole. The results presented in this study were not dependent on the orthogonalization (see SOM). Plots in fig. 2, 3 and 4 were thresholded at $p<0.001$ and restricted to cluster sizes of a minimum of 5 voxels.

Results

Encoding of model based and model free valuations

The ability to test whether the brain generates a System 1 and System 2 value signal requires a task that satisfies two conditions. First, to discriminate between brain activations encoding either value signal, there must be ample opportunity for these valuations to differ. Second, to identify the parameters that capture the subjectivity in either valuation principles, subjects' choices need to follow one principle a significant number of trials, while being consistent with the other one in other trials.

Here, we present evidence that our experimental paradigm satisfied these two conditions. First, while the valuations eventually converged (also evident from Figure 1c), the overall correlation of the estimated System 1 and System 2 values was sufficiently low: r=0.51+-0.05. . The habitual System 1 is mainly characterized in this task by its use of a fixed learning rate (signifying the possibility of a changing environment) as well as a lack of knowledge of the structure of the task leading it to only rely on the direct experienced feedback (i.e. whether a choice was rewarded, punished or had no effect). In contrast the System 2 learner assumes that no changes happen through a session (and thus can integrate over all information) and understands the task well enough to be able to update the value of choices not taken (counterfactuals) based on the feedback. Thus the two systems can lead to markedly different behavior with System 2 converging to the correct values faster than System 1, at the cost of more complex computations (potentially leading to errors in updating (Daw et al., 2005)).

Second, subjects' choices reflected both valuation approaches a sufficient number of times (see Figure 1c). Cross-sectionally, we recorded significant ($p<0.001$) positive correlation between the fits of the two models (see Figure 1b). Subjects evidently differ more in the extent to which they paid attention or were engaged in random exploration, rather than in the extent to which they are predominantly System 1 or System 2 users.

Functional MRI shows activation of both value systems in PFC

We next examined whether there was evidence for separate encoding of System 1 and System 2 value signals. We found separate activations in medial prefrontal cortex (mPFC) that correlated significantly ($p<0.001$ uncorrected, minimum volume 5 voxels) with the System 1 value (Figure 2a; SOM) and the System 2 value of the game (Figure 2b; SOM). There is substantial overlap in these activated areas; no sub-region emerged significantly ($p>0.01$) in the contrast between activation to System 1 value and to System 2 values. Overlap disappeared, however, after applying a stricter criterion to determine regions of significant activation ($p<0.0001$).

Furthermore, subjects whose choices tended to be more in line with System 1 learning did not display relatively stronger activation in mPFC correlating with System 1 value, suggesting that the imaging results are not an artifact of group aggregation (see SOM).

<INSERT FIGURE 2 AROUND HERE>

We next examined value update signals, or *prediction error* signals. It has been shown repeatedly (McClure et al., 2003; O'Doherty et al., 2003) that such signals accompany valuation signals. They provide the crucial input to improving valuation accuracy. We set out to identify and localize activation correlating with the prediction errors for the two learning approaches in our task. In sub-cortical striatal areas (putamen, caudate head, ventral striatum) we recorded phasic activation that correlated ($p<0.001$ uncorrected, minimum volume 5 voxels) only with the prediction error from the System 1 approach (change in RL value of the chosen strategy only, Figure 3a). No subcortical striatal activity correlated with the update of the System 2 (Bayesian) value of the game. The latter update significantly activated left Brodmann area (BA) 10 (Figure 3b), and right inferior frontal gyrus (IFG; for a full list of activations see SOM).

<INSERT FIGURE 3 AROUND HERE>

Pre-Frontal activity precedes subjects' choices

We also found activations that predicted subjects' choices. We found separate activations in mPFC that preceded, 0-4 seconds (adjusted for hemodynamic response delay) before appearance of the choice screen, (i) an optimal Bayesian choice (Figure 4a), (ii) an RL-optimal choice (Figures 4b). Given an average response time of the order of 2 seconds, this activation preceded the actual choice by 4 seconds on average. In Figure 4b, activation in anterior cingulate cortex and globus pallidus correlates negatively, and hence, precedes *deviations* from RL-optimal choice.

To gain further understanding of the influence of the different model components on the decision process we performed a simple psychophysical interaction analysis. PPI is one way to examine the influence of one brain area on another (granger causality and DCM are more advanced versions, see (Valdes-Sosa et al., 2011)), by studying the effect some 'external' variable has on the correlation between two brain areas. As many studies have found that the

ventral striatum is encoding a reinforcement learning error signal, we were interested in how that influences areas in vmPFC.

. We tested whether the correlation between Ventral Striatum and vmPFC was stronger for rounds where subjects performed a choice according to the System 1 model (thus implying a causal relationship). Specifically we used the average activity in the region of Left Ventral Striatum for which activity was found to be significantly correlated with the reward error (see Figure 3a) and looked at the correlation between this region and the activity of the vmPFC for the time period 4 seconds before the onset of the choice screen. We then compared this correlation for rounds where subjects were about to follow System 1 versus rounds where subjects deviated from System 1 and found a significant increase in vmPFC (see table S5 in the Supplementary Online Material for a list of areas, including areas outside PFC), at least part of which (peak activity $p<0.0005$, 15 voxels, uncorrected) overlapped with the area shown in Figure 2 to be encoding the expected reward according to System 1. That is, when subjects are about to perform according to the System 1 strategy the ventral striatum is significantly more correlated with the vmPFC, than when failing to follow such a strategy.

This implies that a strong correlation between Ventral Striatum and prefrontal cortex can prevent subjects from diverging from the correct Model-free strategy, in turn supporting a link between the computation of the Model-free error and Model-free value. Given the strong connectivity from VS to vmPFC/OFC (see e.g., (Haber and Knutson, 2009) for a recent review), such an influence is feasible but we do not know of any studies that have found behavioral correlates of such influence. Also we do want to emphasize that this is merely a correlational result, hence we can not rule out other mechanisms for this effect.

No areas were found where activation correlated significantly (at $p<0.001$ uncorrected, with minimum of 5 voxels) with the absolute difference between the System 1 value of a strategy and the System 2 value of the game. The absence of such a signal of conflict suggests that arbitration between the two systems is not necessarily triggered by competing output (Kerns et al., 2004).

<INSERT FIGURE 4 AROUND HERE>

Relative uncertainty has been suggested as an alternative way to arbitrate (Daw et al., 2005). As measures of the uncertainty for System 2 we used the entropy of the distribution (after taking into account uncertainty in the estimation of this property, see Methods). Behavior was

consistent with the hypothesis that arbitration increased with uncertainty: the full System 2 entropy was larger for rounds where subjects deviated, than for when their choice was guided in accordance with System 2 (significant at $p<0.05$ for 15 out of 17 subjects, unpaired two-tailed t-test with unequal variance, one subject was excluded from this analysis for having too few deviations from System 2).

Regarding deviations from System 1 optimal strategy, we discovered significant predictability based on an estimate of the uncertainty of the System 1 value, namely, an exponentially smoothed average of past squared prediction errors (unpaired two-tailed t-test across subjects, significant at $p<0.005$). This further corroborated the idea that valuations were arbitrated based on uncertainty.

To study this further we created a separate GLM that included these two Uncertainty variables as regressors, as well as the previously specified block regressors. However no brain areas showed enough activity to pass our significance level ($p<0.005$, more than 4 voxels).

Discussion

The presence of two concurrent signals in PFC, one correlating with the System 1 value corresponding to a reflexive, reinforcement learning approach, and another one correlating with the System 2 value of the game representing a reflective, Bayesian learning approach, provides the first unequivocal neurobiological evidence in support of the dual systems approach. The location of the value signals accords with prior findings (Padoa-Schioppa and Assad, 2006; Plassmann et al., 2007; Tobler et al., 2007). Corresponding to the two value signals, we reported two learning signals at outcome presentation. They are (i) the prediction error of the System 1 value of the chosen strategy, and (ii) the update of the System 2 value of the game. The former is located in striatal areas, which suggests that it corresponds to the usual reward prediction error generated by the dopaminergic system (McClure et al., 2003). The signal for the update of the System 2 value of the game is located in frontal cortical areas, in particular, left BA 10. This area has been associated with tasks that, like Bayesian learning, require evaluation of relationships, in particular, deductive reasoning (Monti et al., 2007).

When two different value signals are generated, a resolution mechanism is needed. Two mechanisms have been proposed, one based on conflict (Kerns et al., 2004) and one based on uncertainty of the value signals (Daw et al., 2005). We did not find any support for such a conflict signal (the absolute difference between the System 1 and 2 valuations); engagement of

anterior cingulate cortex (Kerns et al., 2004) and globus pallidus (Yoshida and Tanaka, 2009) as a precursor of deviations of choice from System 1 optimal indirectly suggests adjudication across System 1 valuations through error correction. Consistent with recent theorizing that adjudication need depends on uncertainty, we found that subjects tended to deviate from the System 1 and System 2 optimal strategies when valuation uncertainty was highest.

A few previous papers have examined the issue of model-free versus model-based learning using fMRI. Hampton et al. (Hampton et al., 2006) used a simple reversal learning task and found activity in vmPFC that was better described by a state-based (system 2) learner than a reinforcement (system1) learner, however no analysis of the converse was done. This is compatible with our findings of both systems being present in slightly different area, with parts of vmPFC being better activated by either model.

A very recent paper (Gläscher et al., 2010) also attempted to find neural correlates of the two value systems. While they do not report any value activation for either system, they did find prediction errors for both systems in ventral striatum and lateral PFC for system 1 and 2 respectively, similar to the results we here present for the prediction errors and updating of system 1 and 2 respectively. Our ability to disentangle the value signals can therefore be seen as encompassing the work by Gläscher et al. while extending it by finding the expected values.

While our study was inspired by the work by Daw et al. 2005, our task means we deviate slightly from their approach. For the setup imagined by Daw et al. the main difference between the model free and model based learner is how each builds its model of the world, either updating the values sequentially or building up a transition matrix for the possible states in the system. In our task the main difference between the two systems is in the assumptions underlying the learning, none or fully informed. No assumptions means that there is no knowledge about the structure of the task, which could change without warning, while fully informed means the model knows everything we have told the subject. Specifically this leads to a difference in the treatment of the information, either a steeply discounting model free system versus an integrator that treats all rounds equally. However it is straight forward to see the model-based learner in our task as estimating the probabilities of transition e.g. from a choice to reward states, and acting upon these probabilities. Thus while our experiment was designed to frame the question in a slightly different way than Daw. et al, the analysis can be shown to be

congruent. Despite this difference in interpretation we believe we are true to the spirit of the proposal by Daw et al. and future studies will have to expand upon these ideas.

The role of counter-factuals has been studied in a few recent papers, including by Coricelli et al. (Coricelli et al., 2005), who found encoding of the counterfactual in medial OFC upon revelation. However in our task the counter-factuals are perfectly negatively correlated with the rewards received in the task, making it impossible to distinguish the two.

Our findings may seem to contradict recent reports of a single valuation signal in intertemporal choice tasks (Kable and Glimcher, 2007) or tasks contrasting known (pure risk) and unknown (ambiguity) probabilities (Levy et al., 2010). However, our experiment concerns a situation where, true to the spirit of dual-system theory, the optimal choice can simultaneously be computed based on two different principles (in our case, RL and Bayesian learning). This contrasts with contexts where the values of two options conflict, yet a *single* over-arching valuation principle exists which explains choice. In intertemporal choice tasks, the hyperbolic discounting model provides the unifying valuation principle (Kable and Glimcher, 2007), while alpha-maxmin choice theory explains values and choices in comparisons between pure-risk and ambiguous gambles (Levy et al., 2010).

In summary, our findings constitute neurobiological evidence in favor of dual system theory of decision making. In accordance with recent neuroeconomic evidence, we discovered distinct subjective value signals and prediction error signals corresponding to System 1 and System 2 processes. Only, here we reported *two*, often divergent valuation signals for the *same* choice situation. One was based on a fast way to learn value (RL); the other required more sophisticated thinking, originating in Bayesian updating. Adjudication appeared to happen when uncertainty was highest, consistent with recent theoretical modeling (Daw et al., 2005).

Acknowledgements

The authors would like to thank Peter Dayan for comments on an earlier version of the manuscript. UB was partly supported by the Gatsby Foundation, SQ and PB by grant SES-0527491 from the U.S. National Science Foundation, and PB by the Swiss Finance Institute.

References

Bossaerts, P., Preuschoff, K., Hsu, M., 2008. The neurobiological foundations of valuation in human decision-making under uncertainty. In: Glimcher, P.W., Camerer, C.F., Fehr, E.,

Poldrack, R.A. (Eds.), Neuroeconomics: Decision Making and the Brain. Academic Press, London.
Brainard, D., 1997. The Psychophysics Toolbox. Spatial Vision 10, 433-436.
Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8, 1704-1711.
Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876-879.
De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, Biases, and Rational Decision-Making in the Human Brain. . Science, 684-687.
De Neys, W., Glumicic, T., 2008. Conflict monitoring in dual process theories of thinking. Cognition 106, 1248-1299.
Evans, J.S.B.T., 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. Annual Review of Psychology 59, 255-278.
Fermin, A., Yoshida, T., Ito, M., Yoshimoto, J., Doya, K., 2010. Evidence for model-based action planning in a sequential finger movement task. Journal of motor behavior 42, 371-379.
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron 66, 585-595.
Haber, S.N., Knutson, B., 2009. The Reward Circuit: Linking Primate Anatomy and Human Imaging. Neuropsychopharmacology 35, 4-26.
Hampton, A.N., Bossaerts, P., O'Doherty, J.P., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26, 8360-8367.
Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., Camerer, C.F., 2005. Neural systems responding to degrees of uncertainty in human decision making. Science 310, 1680-1683.
Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10, 1625-1633.
Kahneman, D., Frederick, S., 2002. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In: Gilovich, T., Griffin, D.W., Kahneman, D. (Eds.), Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press, pp. 49-81.
Kahneman, D., Frederick, S., 2007. Frames and brains: elicitation and control of response tendencies. Trends in Cognitive Sciences 11, 45-46.
Kerns, J.G., Cohen, J.D., MacDonald, A.W., III, , Cho, R.Y., Stenger, V.A., Carter, C.S., 2004. Anterior Cingulate Conflict Monitoring and Adjustments in Control. Science, 1023-1026.
Levy, I., Snell, J., Nelson, A.J., Rustichini, A., Glimcher, P.W., 2010. Neural Representation of Subjective Value Under Risk and Ambiguity. Journal of Neurophysiology 103, 1036-1047.
McClure, S.M., Berns, G.S., Montague, P.R., 2003. Temporal prediction errors in a passive learning task activate human striatum. Neuron 38, 339-346.
McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D., 2004. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. Science 306, 503-507.
Monti, M.M., Osherson, D.N., Martinez, M.J., Parsons, L.M., 2007. Functional neuroanatomy of deductive inference: A language-independent distributed network. NeuroImage 37, 1005-1016.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 38, 329-337.
Padoa-Schioppa, C., Assad, J.A., 2006. Neurons in the orbitofrontal cortex encode economic value. Nature 441, 223-226.
Plassmann, H., O'Doherty, J., Rangel, A., 2007. Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. J. Neurosci. 27, 9984-9988.

Platt, M.L., Glimcher, P.W., 1999. Neural correlates of decision variables in parietal cortex. Nature 400, 233-238.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical conditioning II: Current research and theory. Appleton-Century-Crofts, New York, pp. 64-99.
Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychological Bulletin, 3-22.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Tobler, P.N., O'Doherty, J.P., Dolan, R.J., Schultz, W., 2007. Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol 97, 1621-1632.
Tremblay, L., Schultz, W., 1999. Relative reward preference in primate orbitofrontal cortex. Nature 398, 704-708.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124-1131.
Tzieropoulos, H.n., Grave De Peralta, R., Bossaerts, P., Gonzalez Andino, S.L., 2011. The impact of disappointment in decision making: Inter-individual differences and electrical neuroimaging. Frontiers in Human Neuroscience 4.
Valdes-Sosa, P.A., Roebroeck, A., Daunizeau, J., Friston, K., 2011. Effective connectivity: Influence, causality and biophysical modeling. Neuroimage In Press, Corrected Proof.
Yoshida, A., Tanaka, M., 2009. Enhanced Modulation of Neuronal Activity during Antisaccades in the Primate Globus Pallidus. Cereb. Cortex 19, 206-217.

Figure 1: Experimental design and fit of subjective valuation models.

A. Subjects were presented with images of 3 doors and had to rank them by pressing buttons 1, 2 and 3 in the corresponding order (Choice Screen). Responses were self paced. After 5-7 seconds of fixation the correct door would be indicated and the results screen would be displayed for 1.5 seconds showing a gain, no change or a loss (Results Screen) dependent on whether subject had chosen that door as their 1st, 2nd or 3rd preference. After another 5-7 seconds of fixation the next round would begin.

B. The log-likelihood of the RL (System 1) versus that of the Bayesian (System 2) model. The performance of the two models is highly correlated ($p<0.001$).

C. Classification of subjects' choices (first ranked door) over time as Bayesian (green circle) or RL (blue cross) optimal. Note that the two models will agree in certain rounds and that the first trial in each session is not classified.

Figure 2: System 1 and System 2 values are concurrently encoded in mPFC. Saggital slice (x=-3) and coronal slice (y=53) showing the BOLD activation correlated with the System 1 value (A) and System 2 value (B) at the time of the pre-results screen. Threshold was set at $p<0.001$ with a minimum of 5 voxels in each cluster.

Figure 3: System 1 errors are encoded in striatal areas while updates of the System 2 value correlate with activation in BA 10. A) Saggital slice (x=-3) and coronal slice (y=5) showing the activation pattern that significantly correlated with the error in the RL model at the time of the reveal screen. B) Saggital slice (x=26) and coronal slice (y=32) showing the activation pattern that significantly correlated with the change in Bayesian value of the game at the time of the reveal screen. Threshold for all activations was set at $p<0.001$ with a minimum of 5 voxels in each cluster.

Figure 4: Activations in mPFC precede choices guided by either System 1 or System 2. A) Saggital slice (x=-4) and coronal slice (y=1) showing activations predicting choice consistent (yellow; red) or inconsistent (blue) with System 1, 0-4 seconds before display of the decision screen (HRF corrected). B) Saggital slice (x=-4) and coronal slice (y=1) showing increasing activation for a correct (relative to the System 2) decision 0-4 seconds before the decision screen is presented (corrected for HRF). A single area in prefrontal cortex showed activation. Threshold for all activations was set at $p<0.001$ with a minimum of 5 voxels in each cluster.